IE 529
Stats of Big Data and Clustering
Computations: 2
What to submit: DUE DATE– WEDNESDAY, DEC. 20th

For this computational assignment, you should submit a short and concise report of your findings, including the following.

1. Include a scatter plot for each set of data, as is, i.e., with no clusters identified yet.

2. Submit your pseudocodes for the Lloyds/K-means algorithm, the GreedyKcenters and Single Swap algorithms, and the Spectral Clustering algorithm, each on a separate sheet of paper. State clearly whether you are considering normalized or unnormalized Spectral Clustering.

3. For your K-means algorithm by itself, and for your Spectral Clustering algorithm, present the *output* (clustering) results for BOTH sets of test data given to you (see below).

4. Evaluate the results for a range of K values (number of clusters) from K = 2, to K = 11 (but not all of these; try 3 different values, then try to pick 2-3 more that will help you find the best cluster outcome). When implementing the K-means algorithm (by itself and from your Spectral Clustering algorithm), you should try multiple initializations (say 5 to 10) and then select the 'best' result, for each K value. State how many different initializations you tried and why.

   - The *output* should consist of a scatter plot of the data with the clusters differentiated by color, and with the centroids highlighted in a different color, size and shape than the corresponding cluster members.
   - A plot of the distance metric, D, versus, K, for the best outcomes found for each of the K values you considered, for both K-means and Spectral approaches.
   - Using the two best results from your K-means, **that is, two different K-values that seem to give the best clustering result**, use your GreedyKcenters algorithm to find an initialization and re-run K-means with this initialization. Describe how this does or does not improve your results.

5. For the K-means algorithm, describe how you tested for convergence and why you choose this as a test.

6. Compare the results you get using your spectral clustering algorithm to the results obtained with your K-means algorithms for each of the data sets; you should compare the objective function values (D) and discuss qualitatively how well you think each clustering method worked (i.e., base this comparison on the results given in 4).

7. Discuss how many "natural" clusters you think each data set has, and why.

8. Provide a short analysis of the computational effort you found was required for each of the algorithms. State what measure you used to evaluate computational effort (e.g., running time or flops, or ??), and how you observed this. Briefly discuss how these computational efforts compare to your expectations.

9. Write a brief summary paragraph of your findings, that is, briefly summarize your KEY findings.

10. Include your source code in an Appendix, namely .m files (for Matlab routines) or similar for Python or other packages.

11. Please submit the report as a pdf file, with additional attachments as necessary.